The Pennsylvania State University
The Graduate School

USING ANTS TO FIND COMMUNITIES IN COMPLEX NETWORKS

A Thesis in
Computer Science
by
Mohammad Adi

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

May 2014

The thesis of Mohammad Adi was reviewed and approved* by the following:

Thang N. Bui
Associate Professor of Computer Science
Chair, Mathematics and Computer Science Programs
Thesis Advisor

*Signatures are on file in the Gradute School.

Abstract

Many systems arising in different fields can be described as complex networks, a collection of nodes and edges. An interesting property of these networks is the presence of communities (or clusters), which represents a subset of nodes within the network such that the connections within these nodes are denser than the connections with the rest of the network. In this thesis, we give an ant-based algorithm for finding communities in complex networks. Ants are used to identify edges which are used to assign the nodes into different clusters. Tests on various synthetic and real-world networks show that the algorithm is able to extract the community structure very well and performs well against other algorithms.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I am highly grateful to Dr. Thang N. Bui, my thesis advisor, for his guidance and patience throughout the whole work of this thesis. Also, I would like to thank the member's of the thesis committee, ¡insert names¿, for their valuable feedback.

Chapter 1

Introduction

Complex networks are extensively used to model various real-world systems such as social networks, technological (Internet and World Wide Web) networks, biological networks etc. These networks are modeled as graphs where nodes represent the objects in the system and edges represent the relationship among these objects. For example, in a social network, nodes can represent people and two nodes are connected by a link if they are friends with each other.

These networks exhibit distinctive statistical properties. The first property is the "small world effect", which implies that the average distance distance between vertices in a network is short [8]. The second is that the degree distributions follow a power-law [1], and the third one is network transitivity which is the property that two vertices who are both neighbors of the same third vertex, have an increased probabilibty of being neighbors of one another [14].

Another property which appears to be common to such networks is that of community structure (or clustering). While the concept of a community is not strictly defined in the literature as it can be affected by the application domain, one intuitive notion of a community is that it consists of a subset of nodes from the original graph which between them have a higher density of links as compared to their links with the rest of the graph. In this thesis, we describe an ant-based algorithm for detecting communities in graphs.

Over the course of more than a decade, the task of finding communities in networks has received enormous attention from researchers in different fields such as physics,

statistics, computer science etc. As a result, there are currently a vast number of methods which can be used to evaluate the community structure of a network. These methods are described in the next chapter.

Ant algorithms have been previously used to detect communities in graphs [6] [12] [7]. In our approach, we use artificial ants which traverse the graph based solely on local information and deposit pheromone as they travel. This algorithm uses the cumulative pheromone on the edges to build up an initial clustering of the graph. Then a local optimization method is used to reassign the clusters of different nodes based on their degree distribution after which clusters are merged depending on certain rules to obtain the final partitioning of the graph.

The rest of this thesis is organized as follows. Chapter 2 provides more detailed information about the problem statement and covers the previous work done. The ant-based algorithm is described in Chapter 3 and Chapter 4 covers the performance of the algorithm on various synthetic and real-world graphs and compares it to existing algorithms. The conclusion is given in Chapter 5.

Chapter 2

Preliminaries

2.1 Problem Definition

Communities are generally defined to be subsets of vertices which have a high density of links within them. There are various possible definitions of a community and they are divided into mainly three classes: local, global and based on vertex similarity [4] [13]. A more general, quantitative criterion is described in [11] by considering the degree k_i of a node i belonging to a community $S \subset G$, where G is the graph representing the network. The degree of node i can be split as:

$$k_i(S) = k_i^{in}(S) + k_i^{out}(S) \tag{2.1}$$

where $k_i^{in}(S)$ is the number of connections to nodes in its subgraph S and $k_i^{out}(S)$ is the number of connections to nodes outside S. The authors define a community in 2 ways. The subgraph S is a community in the **strong sense** if:

$$k_i^{in}(S) > k_i^{out}(S), \forall i \in S \tag{2.2}$$

The subgraph S is a community in the **weak sense** if:

$$\sum_{i \in S} k_i^{in}(S) > \sum_{i \in S} k_i^{out}(S) \tag{2.3}$$

Even though networks can be directed, undirected, weighted or directed and weighted we concentrate on undirected networks. The problem of community detection can be defined as follows:

Input: An undirected graph $G = (V, E)$ where V represents a set of nodes or vertices and E represents a set of edges or links.
Output: A partition $C = \{C_1, \ldots, C_k\}$ of G into k communities where $C_i \cap C_j = \varnothing, i, j = 1, \ldots, k, i \neq j$ and $C_i \subset V, \forall i$.

2.2 Previous Work

The seminal paper by Girvan and Newman [5], resulted in a lot of research into the area of community detection, especially by physicists. As a result, these days there is a wide variety of community detection algorithms from fields like physics, computer science, statistics etc. Covering all of them is beyond the scope of this work, for a more detailed introduction one can refer the survey by Fortunato [4].

The methods for detecting communities can be broadly classified into hierarchical methods, modularity-based methods, dynamic methods and methods based on statistical properties of the graph [4].

2.2.1 Hierarchical Methods

These type of methods can be further divided into 2 subtypes: divisive hierarchical methods and agglomerative hierarchical methods.

Divisive hierarchical methods start from the complete graph, detect edges that connect different communities based on a certain metric such as edge betweenness [5],

and remove them. Examples of these approaches can be found in [5] [11] [9].

Agglomerative hierachical methods initially consider each node to be in its own community then and merge communities until the whole graph is obtained. Examples can be found in [10] [2] [3].

2.2.2 Modularity-based Methods

Modularity [9] is a metric introduced by Girvan and Newman to evaluate the partitioning of a graph. It is way to quantify the clustering we have obtained in order to determine how good it might be and is a widely adopted quality metric. The idea is that the edge density of the nodes in a cluster should be higher than the expected density of the subgraph whose nodes are connected at random, but with the same degree sequence. This model is called the *null model*.

Using an adjacency matrix representation for the graph, modularity is written as follows:

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(C_i, C_j) \tag{2.4}$$

where A is the adjacency matrix of the graph G, m is the number of edges in the graph, $\frac{k_i k_j}{2m}$ is the expected number of edges between nodes i and j in the null model and δ is the *Kronecker* functions whose value is 1 if i and j are in the same community and 0 otherwise. Since nodes which do not belong in the same cluster don't contribute towards modularity, it can be rewritten as:

$$Q = \sum_{i=1}^{k} \left(\frac{e_i}{m} - \left(\frac{d_i}{2m} \right)^2 \right) \tag{2.5}$$

where k is the number of communities, e_i is the total number of internal links in cluster i and d_i is the sum of the total degrees of nodes in i. So the first term represents the fraction of the total edges that are in a community and the second term

represents the expected value of the fraction of edges in the null model. Values of Q approaching 1 (which is the maximum), indicate strong community structure [9]. In practice, the value usually ranges from 0.3 - 0.7.

References

[1] Albert-Laszlo Barabasi and Reka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[2] V.D. Blondel, J.L. Guillaume, R. Lambiotte, and E.L.J.S. Mech. Fast unfolding of communities in large networks. *J. Stat. Mech*, page P10008, 2008.

[3] Aaron Clauset, M. E. J. Newman, , and Cristopher Moore. Finding community structure in very large networks. *Physical Review E*, pages 1– 6, 2004.

[4] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(35):75 – 174, 2010.

[5] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.

[6] Dongxiao He, Jie Liu, Bo Yang, Yuxiao Huang, Dayou Liu, and Di Jin. An ant-based algorithm with local optimization for community detection in large-scale networks. *CoRR*, abs/1303.4711, 2013.

[7] Di Jin, Dayou Liu, Bo Yang, Carlos Baquero, and Dongxiao He. Ant colony optimization with markov random walk for community detection in graphs. In *Proceedings of the 15th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining - Volume Part II*, PAKDD'11, pages 123–134, Berlin, Heidelberg, 2011. Springer-Verlag.

[8] Stanley Milgram. The Small World Problem. *Psychology Today*, 2:60–67, 1967.

[9] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69(2):026113, February 2004.

[10] M.E.J. Newman. Fast algorithm for detecting community structure in networks. *Physical Review E*, 69, September 2003.

[11] Filippo Radicchi, Claudio Castellano, Federico Cecconi, Vittorio Loreto, and Domenico Parisi. Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(9):2658–2663, 2004.

[12] S. Sadi, S. Oguducu, and A.S. Uyar. An efficient community detection method using parallel clique-finding ants. In *Evolutionary Computation (CEC), 2010 IEEE Congress on*, pages 1–7, 2010.

[13] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

[14] D.J. Watts and S.H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, (393):440–442, 1998.